ISSN: 2664-9640

Research on the Construction of Personal Credit Score Card Model based on Logistic Regression and Random Forest Fusion Model

Liying Yan, Yukai Ma, Yao Ming

School of Management, Minzu University of China, Haidian, Beijing 100080, China

Abstract

This paper discusses the indicators and models of personal credit score cards, aiming to improve the accuracy and efficiency of evaluation. Through the literature review, we compare the traditional and machine-learning, deep learning-based scoring models, and focus on the combined prediction models to cope with the complex economic situations. Data processing was performed using random forest predicting missing data and XGBoost feature weights assisted feature selection. The model constructed a combined model of random forest and logistic regression and a credit score card assisted by the XGBoost model, which proved that the prediction effect of the combined model was better than that of a single model.

Keywords

Personal Credit Score Card; Logistic Regression; Xgboost Model; Random Forest.

1. Introduction

Economic development and consumption upgrading drive the growth of personal credit, but China's credit system has information asymmetry, data island and privacy problems, which affect the accuracy of evaluation.[1]Big data and cloud computing have promoted the reform of the credit investigation model, but the problems of the laws, standards and the independence of the credit investigation agencies still need to be solved. China's credit investigation system is not mature, facing a large population base and data statistics problems. Financial institutions should learn from their experience, innovate evaluation methods, establish advanced models that adapt to the current trend of digitalization and intelligence, and realize the balance between risk control and profit pursuit. The team uses big data, combines logical regression and random forest model to build scoring cards, improve the efficiency and accuracy of evaluation, make up for the shortage of existing models, enrich credit evaluation research, and promote the development of bank fintech.

2. Summary of Studies on Personal Credit Score Index Systems and Models

The development of personal credit evaluation system has gone through many international and domestic stages, and has now formed a diversified evaluation index system.[2]In 1941, David Durand proposed a personal credit evaluation index system including nine factors, including age and gender. In 1956, Bill Faiif and Earl Isaac created the personal credit Risk FICO score, including credit repayment performance, credit history service life, etc. In 1979, FFIEC issued the CAMELS credit rating system, including five core indicators including capital adequacy ratio and management ability. Domestic commercial banks mainly evaluate customer credit based on their personal basic information, economic status and historical credit records.

Zhang Chen and Wan Xiangyu suggested building a personal credit evaluation system based on big data, including six dimensions of personal characteristics and economic ability.

The personal credit evaluation model covers economic statistics, statistics and econometrics, artificial intelligence and big data methods. Financial institutions mainly use statistical and econometrics methods, while introducing machine learning models such as random forests.[3]The development of the model tends to diversify, combining artificial intelligence and integrated learning methods to improve prediction accuracy and stability, and the combined prediction methods to improve accuracy by fusion of various technologies.

3. Data Preprocessing

3.1. Data Source

From the more than 1.2 million loan records of a credit platform, 800,000 were selected as training sets, and the remaining 200,000 were divided into test sets A and test sets B, and some fields were desensitized. Based on the data timeliness, observation period and sample size, 8 million pieces of data and 46 features were finally selected for analysis and modeling, covering the loan amount, grade, historical credit record and other information.

3.2. Data Cleaning and Replenishment

In the data cleaning stage, the samples with a large proportion of characteristic missing values are excluded, and the post-loan characteristics of the leaked default information are deleted, and the unreasonable data is eliminated. For the incomplete data, random forest was used to predict the employment years with an accuracy of 98.09%. The complete data had little influence on the model. Numerical deletions are filled with the median, and categories with the means.

3.3. Data Normalization

Due to the different feature scales of known data, this paper uses zero mean standardization to normalize the data by reducing the dependence of the model on the specific feature scale and improving the generalization ability of the model on different data distribution.

3.4. Selection of Variables

During the data preprocessing phase, Abexception filtering, missing data interpolation, feature binning and WOE coding, 10 features were selected for the final model, With isDefault as the dependent variable, Model analysis was performed using interestRate, term_bin, issue Date_bin, homeOwnership, ficoRangeLow, VerificationStatus, loanAmnt, applicationType, grade9 original variables and new feature openAcc_totalAcc.

4. Model Optimization

4.1. Model Selection

Although XGBoost and neural networks have better performance, we prefer to use logistic regression model in risk control modeling, because of its simplicity and strong interpretability, which facilitates problem diagnosis and feature coefficient business evaluation. The following is the decision process generated by XGBoost.

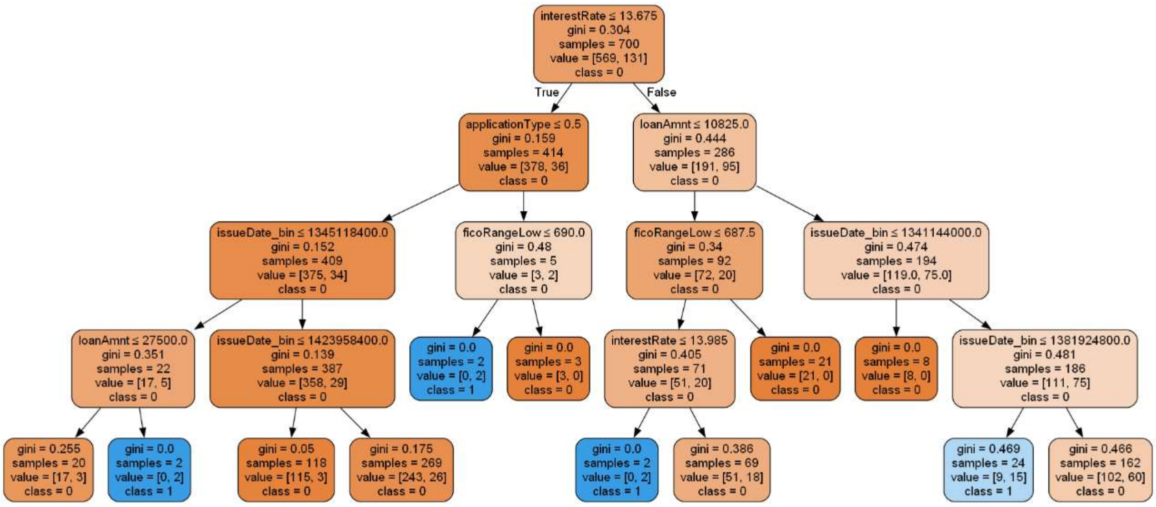


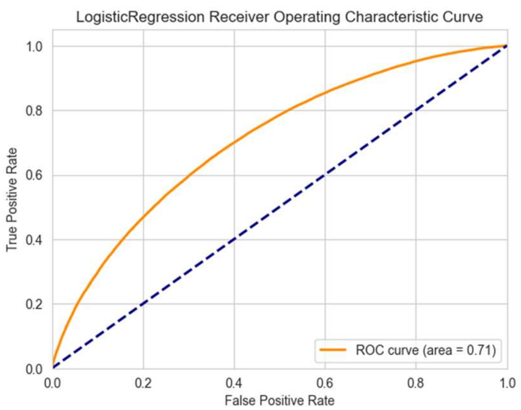
Figure 1. XGBoost decision tree

4.2. Model Evaluation

The credit score card model is evaluated using AUC indicators and KS (Kolmogorov-Smirnov) values to measure the ability of the model to distinguish positive and negative samples to calculate the TPR and FPR at different thresholds. The final KS value of the model is the maximum KS value of each threshold value. Model validation KS value is 0.31 and the model is available.

4.3. Pre-training Model

To ensure that the model learning is sufficient and stable, the data set is divided into the training set and the test set in a 7:3 ratio. Model training features or hyperparameter optimization used hierarchical 5-fold cross validation, in which the necessity of feature crossing was assessed by auxiliary model XGBoost. The comparison of difference determines whether to adjust the features. If the difference is large, the model stability is insufficient or the data set is different, it is necessary to adjust the feature optimization model or reformulate the data set division strategy.



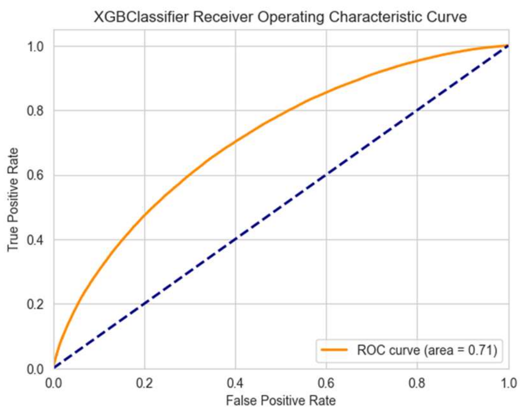


Figure 2. The ROC plot of the logistic regression model **Figure 3.** XGBoost model ROC plot

According to the results of XGBoost performance evaluation results in the final training set, its AUC = 0.713 and KS = 0.308, which showed no significant improvement over the original value of the model. Therefore, the current use of features does not require cross-combination to improve the nonlinear capability of the model.

4.4. Score Card Generation

Set the P according to the business needs0, score0, Rate, PDO several values, and calculate the offset and factor generation scoring function. support score0Set to 600 points, P0 to 20:1, to 50 and rate to 2.

index	feature	interval	score
0	interestRate	(-inf, 7.96]	147
1	interestRate	(7.96, 10.41]	74
2	interestRate	(10.41, 12.59]	32
3	interestRate	(12.59, 13.99]	-2
4	interestRate	(13.99, 15.99]	-23
5	interestRate	(15.99, 17.8]	-48
6	interestRate	(17.8, 20.9]	-72
7	interestRate	(20.9, inf]	-108
8	term_bin	(-inf, 5.0]	27
9	term_bin	(5.0, inf]	-67
10	issueDate_bin	(-inf, 1391212800.0]	32
11	issueDate_bin	(1391212800.0, 1412121600.0]	12
12	issueDate_bin	(1412121600.0, 1459468800.0]	-1
13	issueDate_bin	(1459468800.0, 1506816000.0]	-26
14	issueDate_bin	(1506816000.0, inf]	14
15	homeOwnership	(-inf, 1.0]	19
16	homeOwnership	(1.0, 2.0]	-20
17	homeOwnership	(2.0, inf]	-5
18	ficoRangeLow	(-inf, 675.0]	-34
19	ficoRangeLow	(675.0, 685.0]	-21
20	ficoRangeLow	(685.0, 705.0]	-2
21	ficoRangeLow	(705.0, 720.0]	23
22	ficoRangeLow	(720.0, 735.0]	43
23	ficoRangeLow	(735.0, inf]	82
24	verificationStatus	(-inf, 1.0]	38
25	verificationStatus	(1.0, 2.0]	-6
26	verificationStatus	(2.0, inf]	-23
27	loanAmnt	(-inf, 3525.0]	39
28	loanAmnt	(3525.0, 9500.0]	23
29	loanAmnt	(9500.0, 10025.0]	8
30	loanAmnt	(10025.0, 15025.0]	-5
31	loanAmnt	(15025.0, 28025.0]	-16
32	loanAmnt	(28025.0, inf]	-26
33	openAcc_totalAcc	(-inf, 0.31645569620253167]	19
34	openAcc_totalAcc	(0.31645569620253167, 0.4117647058823529]	13
35	openAcc_totalAcc	(0.4117647058823529, 0.46938775510204084]	7
36	openAcc_totalAcc	(0.46938775510204084, 0.5172413793103449]	2
37	openAcc_totalAcc	(0.5172413793103449, 0.5671641791044776]	-1
38	openAcc_totalAcc	(0.5671641791044776, 0.6071428571428571]	-7
39	openAcc_totalAcc	(0.6071428571428571, 0.6527777777777778]	-12
40	openAcc_totalAcc	(0.6527777777777778, 0.7536231884057971]	-19
41	openAcc_totalAcc	(0.7536231884057971, inf]	-28
42	applicationType	(-inf, inf]	0
43	grade	(-inf, 2.0]	138
44	grade	(2.0, 3.0]	50
45	grade	(3.0, 4.0]	-16
46	grade	(4.0, 5.0]	-57
47	grade	(5.0, inf]	-104
48	basic_score	nan	822

Figure 4. Binning results and WOE values

index	feature	interval	woe
0	interestRate	(-inf, 7.96]	-1.43983
1	interestRate	(7.96, 10.41]	-0.722417
2	interestRate	(10.41, 12.59]	-0.310451
3	interestRate	(12.59, 13.99]	0.024275
4	interestRate	(13.99, 15.99]	0.226419
5	interestRate	(15.99, 17.8]	0.473895
6	interestRate	(17.8, 20.9]	0.709835
7	interestRate	(20.9, inf]	1.062366
8	term_bin	(-inf, 5.0]	-0.268618
9	term_bin	(5.0, inf]	0.652008
10	issueDate_bin	(-inf, 1391212800.0]	-0.31068
11	issueDate_bin	(1391212800.0, 1412121600.0]	-0.113416
12	issueDate_bin	(1412121600.0, 1459468800.0]	0.013479
13	issueDate_bin	(1459468800.0, 1506816000.0]	0.259284
14	issueDate_bin	(1506816000.0, inf]	-0.136825
15	homeOwnership	(-inf, 1.0]	-0.185441
16	homeOwnership	(1.0, 2.0]	0.19289
17	homeOwnership	(2.0, inf]	0.051096
18	ficoRangeLow	(-inf, 675.0]	0.334991
19	ficoRangeLow	(675.0, 685.0]	0.205601
20	ficoRangeLow	(685.0, 705.0]	0.015302
21	ficoRangeLow	(705.0, 720.0]	-0.226416
22	ficoRangeLow	(720.0, 735.0]	-0.416307
23	ficoRangeLow	(735.0, inf]	-0.798926
24	verificationStatus	(-inf, 1.0]	-0.36722
25	verificationStatus	(1.0, 2.0]	0.060872
26	verificationStatus	(2.0, inf]	0.224884
27	loanAmnt	(-inf, 3525.0]	-0.379938
28	loanAmnt	(3525.0, 9500.0]	-0.228876
29	loanAmnt	(9500.0, 10025.0]	-0.080889
30	loanAmnt	(10025.0, 15025.0]	0.052078
31	loanAmnt	(15025.0, 28025.0]	0.157041
32	loanAmnt	(28025.0, inf]	0.258799
33	openAcc_totalAcc	(-inf, 0.31645569620253167]	-0.190157
34	openAcc_totalAcc	(0.31645569620253167, 0.4117647058823529]	-0.127958
35	openAcc_totalAcc	(0.4117647058823529, 0.46938775510204084]	-0.070446
36	openAcc_totalAcc	(0.46938775510204084, 0.5172413793103449]	-0.021408
37	openAcc_totalAcc	(0.5172413793103449, 0.5671641791044776]	0.007758
38	openAcc_totalAcc	(0.5671641791044776, 0.6071428571428571]	0.064406
39	openAcc_totalAcc	(0.6071428571428571, 0.6527777777777778]	0.118083
40	openAcc_totalAcc	(0.6527777777777778, 0.7536231884057971]	0.187507
41	openAcc_totalAcc	(0.7536231884057971, inf]	0.272294
42	applicationType	(-inf, inf]	0
43	grade	(-inf, 2.0]	-1.355566
44	grade	(2.0, 3.0]	-0.485412
45	grade	(3.0, 4.0]	0.152694
46	grade	(4.0, 5.0]	0.560321
47	grade	(5.0, inf]	1.018406

1Figure 5. A scoring card

4.5. Verification of the Validity

The mapping function is made according to the generated scoring card, and the incoming customer samples are automatically calculated. Finally, compare the difference between good and bad customers. If the score of a bad customer is less than that of good customers, it indicates that the scoring card is valid.

Table 1. Assessment of the model prediction results

Average good customer score:	1033.09
Mean bad customer score:	930.66
pvalue:	0.0

The results show that the average score of bad customers is less than that of good customers, and the difference is significant, and the score card is valid.

5. Summary

In this paper, a personal credit score card model combining random forest, time logistic regression and XGBoost is constructed to calculate the scores of variables in different bins and add them to get the final score. The model has practical application value, but the data acquisition and privacy protection limit its application scope, and the single lending platform data is used to build the model, and the statistical evaluation index is limited, which needs further verification.

References

[1] Zhang Junli, Guo Shuangyan, Ren Cuiping, Ma Qian. Study on personal credit score card model based on logistic regression [J]. Modern Information Technology, 202,8 (05): 12-16.

[2] Durand D.Risk Elements in consumer Installment financing[J].National Bureau of Economic Research, 1941:60~72.

[3] Cui Boyu. A Credit Risk Assessment Model and Applied Research Based on Cost-Sensitive and Incremental Learning [D]. Southwestern University of Finance and Economics, 2022.